

DOCUMENT RESUME

ED 393 915

TM 024 923

AUTHOR Wheeler, Patricia H.
TITLE Functional-Level Testing: A "Must" for Valid and Accurate Assessment Results. EREAPA Publication Series No. 95-2.
INSTITUTION EREAPA Associates, Livermore, CA.
PUB DATE 95
NOTE 18p.
AVAILABLE FROM EREAPA Associates, 2840 Waverley Way, Livermore, CA 94550-1740 (\$6).
PUB TYPE Guides - Non-Classroom Use (055)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Adaptive Testing; *Difficulty Level; *Educational Assessment; Performance; *Scores; *Scoring; *Testing Problems; *Test Results; Test Validity
IDENTIFIERS Accuracy; *Functional Level Testing

ABSTRACT

When individuals are given tests that are too hard or too easy, the resulting scores are likely to be poor estimates of their performance. To get valid and accurate test scores that provide meaningful results, one should use functional-level testing (FLT). FLT is the practice of administering to an individual a version of a test with a difficulty level that best matches the test taker's current performance. The often-confused terms "in-level" and "out-of-level" testing refer to the assessment itself and the group for which it was designed, but FLT refers to the individual being assessed and to matching the difficulty level of the test to the individual's expected level of performance. To implement FLT, the teacher must first determine the expected level of performance and then ensure that there is test material for all levels. There are challenges to administering the FLT including the careful administration and scoring of the tests and the difficulty in explaining results so that test takers feel they were treated fairly. Some issues in scoring and reporting results are discussed. (Contains six tables and three references.) (SLD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *



U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.
☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

PATRICIA H. WHEELER

EREAPA
Publication
Series
No. 95-2

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

FUNCTIONAL-LEVEL TESTING:

A *MUST* FOR VALID AND

ACCURATE ASSESSMENT RESULTS

Patricia H. Wheeler, M.B.A., Ph.D.

FUNCTIONAL-LEVEL TESTING:

A *MUST* FOR VALID AND

ACCURATE ASSESSMENT RESULTS

Patricia H. Wheeler, M.B.A., Ph.D.

EREAPA Associates
2840 Waverley Way
Livermore, California 94550-1740

Copyright © 1995 by Patricia H. Wheeler.

The author expresses her appreciation to Ed Lee, Modesto (CA) City School District, and Charles Woodson, University of California at Berkeley, for their comments and suggested changes an earlier version of this paper, and to Paul K. Wheeler for improving this paper's readability, clarity, and graphics. Final responsibility for the contents rests with the author.

**FUNCTIONAL-LEVEL TESTING:
A MUST FOR VALID AND ACCURATE ASSESSMENT RESULTS**

Patricia H. Wheeler, M.B.A., Ph.D.
EREAPA Associates
Livermore, California

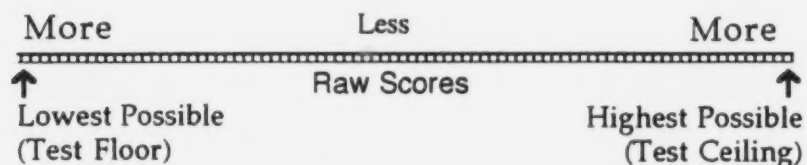
A PRIMER ON FUNCTIONAL-LEVEL TESTING

When individuals are given tests that are too difficult or too easy, the resulting test scores are likely to be poor estimates of their performance. If you want valid and accurate test scores that provide meaningful assessment results, you should use functional-level testing, or FLT. Before exploring the advantages and disadvantages of FLT and discussing implementation issues and data interpretation, we must discuss some basic concepts in measurement so you can better understand the importance of adopting FLT practices in your assessment program.

Every test has a ceiling and a floor. The *ceiling* represents the highest level of performance that a test can measure, and the *floor*, the lowest level of performance. The distance between the ceiling and floor is the range of performance that can be estimated by that test. Normally, to have a wide range, a test would need to have not only a large number of items measuring the attribute in question, but also items of varying degrees of difficulty, from very easy to very hard, for the individuals who would take that test.

To obtain a reasonably accurate estimate of an individual's performance, we want to administer a test at a level on which his or her score would be likely to fall in the middle of the score distribution. The closer the individual's score is to the middle, the less error of measurement (i.e., systematic error due to factors inherent in the assessment instrument or process), and the more accurate the estimated performance of that individual.

Error of Measurement



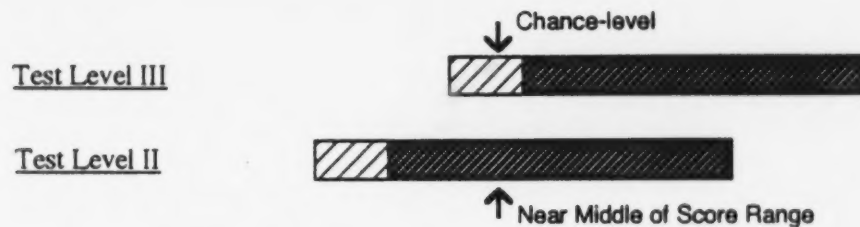
Ingebo (1994), in an analysis of the Portland (OR) Public Schools' district testing program, found that functional-level testing provided more accurate results than grade-level testing.

Instead of having very long tests, publishers typically offer several different levels of a test. The *level* designates "the difficulty for a version of a test that is part of a series of articulated tests (Wheeler and Haertel, 1993, p. 83). The level can reflect a hierarchy of characteristics such as age, grade, number of courses taken, or years of experience.

To illustrate the concept of test level, imagine a mathematical problem solving test that has three levels: Level I for grades 3 and 4, Level II for grades 4 and 5, and Level III for grades 5 and 6. In this example, Level I would have the easier items and Level III the more difficult items from a pool of items for grades 3-6. We could combine all the items from this pool into one test version covering all four grades. It need not be as long as the sum of the number of items on the tests at all three levels. In fact, some of the harder Level I items might be on Level II and the easier Level III items on Levels I and II. But it would most likely be a longer test or, if kept the same length as the tests at each level, may not cover as many mathematical skills.

If a fifth grader were performing poorly in mathematics class and that student were given Level III along with the rest of the class, the student might fare very badly—possibly at chance or lower—and we would not obtain a good estimate of how much mathematics that student understood. The *chance score* for an individual is the number of items attempted by that individual times the probability of responding correctly to a item by random guessing (Wheeler and Haertel, 1993). If the student only answered 20 of the 35 items on the test and there were four choices per item, that individual's chance-level score would be 20 times $1/4$, or 5. The *chance-level score* for a test is based on the assumption that the examinee responds to *all* items, but by random guessing. On a 35-item test with four choices per item, a chance-level score would be 35 times $1/4$, or 9. In the case of this fifth grader, if the student took Level II or possibly even Level I, we might obtain a reasonably accurate estimation of performance in mathematics. Even though Level III was designed for fifth graders and has a reasonable overlap with the content and skills being taught in that student's classroom, a low score tells us nothing about how well the student understands mathematical problem solving.

Any examinees performing at chance-level, near the floor of the test, should be given an easier level of the test. Gabriel and Evans (undated) illustrate this as follows:



Correspondingly, examinees scoring very high, near the ceiling of the test, should be given a more difficult level to obtain a more accurate estimate of their performance.



Functional-level testing (FLT) is the practice of administering a version of a test to an individual with a difficulty level that best matches the test taker's current performance. Terms often confused with FLT are *in-level* and *out-of-level* testing. In-level testing means using a test level that was developed for persons of the same age, grade, or background as the examinees. Out-of-level testing means using a test level that was developed for persons of ages, grades, or backgrounds different from the examinees. One level of an assessment might be designed for medical technicians who have no on-the-job experience, only formal training, while a more difficult level, possibly intended for advanced licensing or promotion, might be designed for medical technicians with five to ten years of laboratory experience beyond formal training. If an inexperienced medical technician were given the easier assessment, that would be in-level testing. If an experienced medical technician took the easier assessment, or the inexperienced technician the more difficult assessment, that would be out-of-level testing.

In-level and out-of-level testing refer to the assessment itself and the group for whom it was designed. FLT refers to the individual being assessed and to matching the difficulty

level of the test to that individual's expected level of performance. Table 1-below lists some of the advantages and disadvantages of FLT. These issues are discussed more fully in the remainder of this paper.

Table 1. Advantages and Disadvantages of Functional-Level Testing.

ADVANTAGES OF FLT	DISADVANTAGES OF FLT
• Yields more accurate estimates of an examinee's performance. • Avoids floor and ceiling effects. • May have higher validity if it is a better match to what the examinee is learning or doing. • Is less frustrating to the examinee and can be a more positive experience and give feelings of success. • Discourages random guessing and capricious test-taking behavior that can be disruptive to other examinees. • Promotes good test-taking skills. • Provides a better estimate not only of current performance, but of change in performance over time.	• May increase the cost of test materials. • Can be difficult to determine which assessment level to give to each examinee. • Can lead to administrative headaches if there are different directions or time limits for each level. • May not be developmentally appropriate for the examinee. • Depending on the use of the results, may not be valid. • May not have been vertically equated in a sound manner and even then only to adjacent levels. • Requires more care in scoring and interpreting. • Harder to explain, as FLT tends to give the appearance of unfairness.

ISSUES WHEN IMPLEMENTING FUNCTIONAL-LEVEL TESTING

As a first step, you must determine each individual's expected level of performance. Usually classroom teachers have sufficient information about their students to make such determinations. But in the case of new students or individuals with whom you are not familiar, it can be challenging to decide which level of assessment to give. Some publishers offer so-called "locator tests" to help identify the test level an individual should take. If you have previous test scores assessing the same or a related attribute of the examinees, you may be able to use those scores to determine which level should be

given. Examinees with very low scores on the earlier test, or who earned chance scores, should take an easier level. Those near the ceiling of the earlier test should be considered for a more difficult test level.

Another pitfall of FLT is that if you find you should give out-of-level tests, they may not be developmentally appropriate to the individuals. For example, you would not want to give a non-reading adult a first-grade reading skills test as the content would be too childish or immature for the adult. You also must be careful not to give even the brightest kindergartner a test for higher elementary grades that has small print and requires a separate answer sheet. Thus, alternative levels should not only match the individuals' functional levels in terms of the difficulty of the skills and content assessed, but should also be developmentally appropriate for them.

When doing FLT and using out-of-level tests, you may lack content validity. That is, the test may not be measuring content and skill areas that are relevant for the intended assessment. If you are assessing how well your students are learning the intricacies of a word-processing package and you give those students who are finding this software challenging an assessment on the basics of keyboards, you may obtain a reasonable estimate of how much they know about keyboards, but you will not know how effective your instruction was because the test is not covering what you are teaching, that is, word processing.

You must take care when ordering test materials to ensure you have test materials for all levels. If you would normally order several different levels of a test for your students or examinees, this may not be a problem. But when the test materials are received, you may need to reassemble the test booklets, answer documents, and manuals of directions for each testing group. When packages of test booklets are opened before the testing session, precautions must be taken not to lose a single test booklet. It is much easier to notice a missing package of test booklets than a single booklet. In addition, you may have to purchase additional testing materials. Even though the total number of test booklets required may be the same for in-level testing or FLT, you may need to buy additional packages of test booklets and answer sheets for other levels as well as technical manuals or norms booklets if they are issued by level.

You can encounter administrative challenges. Different levels of tests may have different directions or time limits. When multiple levels of a test are being given in a classroom

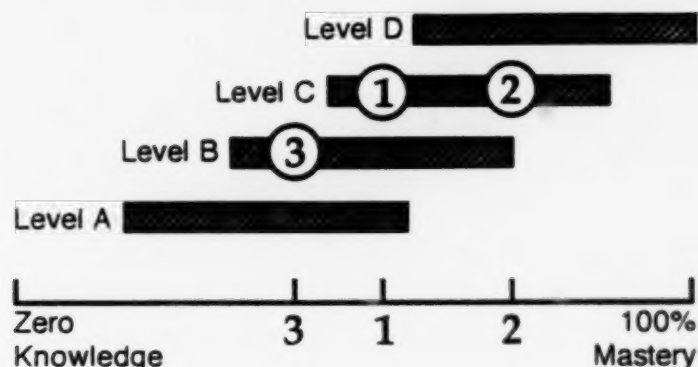
simultaneously, an administrative nightmare can ensue. You must be certain that each examinee has the appropriate test booklet and answer document for the level he or she is to take. Some test publishers have been moving toward having similar time limits and test directions for adjacent test levels, but this is not always the case. When the time limits or test directions differ, the examinees can be split up into separate rooms for the testing, or take the tests on separate days or at separate times.

Furthermore, FLT requires more careful processing of answer sheets and test responses. You must maintain an accurate record of which level (and form) each examinee actually took. It is easy to find yourself using the wrong answer key, scoring rubric, conversion table, or norms. However, current scoring and data analysis software programs should ameliorate this problem.

Finally, you should take care when reporting and explaining assessment results to ensure that people do not feel they were treated unfairly by being given a harder level or that others had an unfair advantage over them because they were given an easier level of the assessment. The fairness issue has to be addressed early in the process, at the time you are determining which level of a test to give an examinee so that you can justify your decisions and address any challenges you might encounter later. Nevertheless, with FLT, you have the option of retesting an examinee with a different level of an assessment.

Remember also that test-taking behavior can be adversely affected if functional-level testing practices are *not* employed. Those examinees confronted with too difficult or too easy a test can become upset, disruptive, depressed, frustrated, bored, and sloppy. The testing can be a negative experience for them. In group testing situations, such behaviors will not only pose a challenge to the person administering the tests, but can also distract the other test takers and negatively impact their performance.

Not only will FLT give you better estimates of an individual's performance, but also better information on changes in performance between the pre-test and post-test, an important consideration for evaluating not just individuals' progress, but also a program's impact. Gabriel and Evans (undated) offer the following example to clarify this issue.



Gabriel and Evans (1986, p. 7) explain this as follows:

Suppose some students are given Level C, but the test is too difficult for them. They may guess on many items and score in the chance range of the test—at point 1. Then let's assume their posttest performance shows improvement and they score at point 2. Their *apparent gain* is the distance between the pretest and posttest (2 minus 1). But suppose these same students were pretested with Level B, which is closer to their functional level, and then, because of growth during the year, they are posttested with Level C, as in the first example, and obtained the same posttest score (point 2). The students' *actual gain* from pretest to posttest is 2 minus 3. In reality, they grew much more during the year than the original test sequence showed.

Before deciding on FLT, you must be certain that the test publisher designed and implemented a sound vertical equating process. Vertical equating is accomplished by having a group of examinees take two levels of the test, half of them taking the easier level first and half of them the harder level first. Then conversion tables are developed so that raw scores on each level convert to scores on a common scale (i.e., scaled scores). Tests are usually vertically equated to adjacent levels, but rarely to non-adjacent levels. If you are using a level that is two or more levels away from the in-level test, the vertical equating results are based on less solid procedures. For example, the first two shown below are directly equated; the third one is indirectly equated through the first two.

Level I $\longleftrightarrow$ Level II

Level II $\longleftrightarrow$ Level III

Level I $\cdots\cdots\cdots\longleftrightarrow$ Level III

DATA INTERPRETATION IN FUNCTIONAL-LEVEL TESTING

This section illustrates the conversion of raw scores and the interpretation of test data in functional-level testing practices. Pages 12-13 show partial norms tables for Levels I and II of the mathematical problem solving test described earlier. Note that there are also two forms of the test available at each level, Form A and Form B. Having alternative, or parallel, forms at each level enables you to retest an examinee at the same level without using an identical form. Page 14 has the *Percentile Rank/Normal Curve Equivalent Conversion Table*. A form is shown on page 15 on which a teacher might record test results for some students. Using the two norms tables and the conversion table on page 14, you can complete the information for this mythical cohort of fifth graders.

For example, Lynn took Level I, Form A in the fall and Level II, Form B in the spring. Since Lynn took Level I, you must start with the first norms table. Find Lynn's raw score under the Form A column and locate the corresponding scaled score. In this case a raw score of 29 on Level I, Form A, has a scaled score of 793. Then go to the norms table for Level II, which includes percentile ranks for grade 5, and look for 793 in the scaled score column. Follow this line across to the column for fall of grade 5 to obtain the percentile rank for Lynn; the percentile rank is 35. Use the conversion table on page 14 to locate the normal curve equivalent (NCE) corresponding to a percentile rank of 35, in this case, 41.9.

Raw Score for level taken —————> Scaled Score

Scaled Score —————> Percentile Rank —————> NCE

Repeat this process for all the students on the test results sheet, and then compare your figures to those on the page 16. For more information on the meaning of these and other types of test scores see Wheeler and Haertel (1993).

CONSIDER THE ALTERNATIVES

Computer-adaptive testing is one technique for moving toward FLT for all examinees. Here the test itself adjusts to the examinees' functional levels, giving each examinee a tailored set of test items. If an examinee answers the first six or seven items of medium difficulty correctly, he or she will be shown more difficult items. However, if he or she misses most of them, easier items will come up on the computer screen. But given that

we are still heavily dependent on assessment instruments that have the level of difficulty built in, even for alternative assessments (e.g., essays, interviews, projects), the flexibility inherent in computer-adaptive testing is not always feasible. When you have an array of assessments or tests of varying levels of difficulty, you should carefully consider the tradeoffs of functional-level testing.

Lastly, you must remember that a test or assessment only provides an estimate of an individual's performance, based on the array of content and skills that the test covers and the nature of the assessment (e.g., difficulty of the content, type of measure, time limits, scoring subjectivity). By testing each individual at a level that matches his or her functional level, you offer them the best opportunity to show what they know and what they can do, and you make the testing a more positive experience by decreasing the possibilities of examinee boredom or frustration when they are confronted with a very easy or very difficult test. To do otherwise would be unfair to the examinees.

Level I (Grades 3-4)						
Raw Score: Form		Scaled Score	Percentile Rank			
A	B		Grade 3		Grade 4	
			Fall	Spring	Fall	Spring
33	35	800	77	62	60	43
	34	799	76	60	58	41
32		798	76	60	57	41
	33	797	76	60	57	41
31		796	74	59	56	39
30		795	72	59	56	39
	32	794	71	57	54	36
29		793	71	57	54	36
	31	792	69	56	53	34
28	30	791	68	54	53	33
	29	790	66	54	52	32
27	28	789	65	53	51	29
26		788	64	52	51	29
25	27	787	63	50	48	28
24	26	786	61	47	45	26
	25	785	60	46	43	25
23		784	58	45	42	23
	24	783	54	43	39	22
22	23	782	53	41	37	22
	22	781	51	40	36	20
21		780	48	37	32	19
	21	779	48	37	32	19
20		778	45	35	29	17
	20	777	45	35	29	17
19		776	43	32	26	16
18	19	775	40	31	24	16
	18	774	36	29	21	15
17		773	36	29	21	15
16	17	772	34	28	20	15
15	16	771	33	26	18	14

Level II (Grades 4-5)						
Raw Score: Form		Scaled Score	Percentile Rank			
A	B		Grade 4		Grade 5	
			Fall	Spring	Fall	Spring
28		800	60	43	42	33
	27	799	58	41	40	33
27	26	798	57	41	40	33
26		797	57	41	40	33
	25	796	56	39	37	32
25		795	56	39	37	31
	24	794	54	36	35	30
24		793	54	36	35	29
	23	792	53	34	33	29
23	22	791	53	33	31	28
	21	790	52	32	30	27
22		789	51	29	27	25
21	20	788	51	29	27	24
	19	787	48	28	26	22
20		786	45	26	23	20
	18	785	43	25	22	18
19		784	42	23	21	17
	17	783	39	22	20	16
18		782	37	22	20	16
17		781	36	20	18	14
	16	780	32	19	17	13
16		779	32	19	16	12
15		778	29	17	14	11
	15	777	29	17	13	10
14		776	26	16	13	10
	14	775	24	16	11	08
13	13	774	21	15	11	08
	12	773	21	15	10	06
12		772	20	15	10	06
11	11	771	18	14	08	05

PERCENTILE RANK/NORMAL CURVE EQUIVALENT CONVERSION TABLE *

<u>Percentile Rank</u>	<u>NCE</u>	<u>Percentile Rank</u>	<u>NCE</u>	<u>Percentile Rank</u>	<u>NCE</u>
1	1.0	34	41.3	67	59.3
2	6.7	35	41.9	68	59.9
3	10.4	36	42.5	69	60.4
4	13.1	37	43.0	70	61.0
5	15.4	38	43.6	71	61.7
6	17.3	39	44.1	72	62.3
7	18.9	40	44.7	73	62.9
8	20.4	41	45.2	74	63.5
9	21.8	42	45.8	75	64.2
10	23.0	43	46.3	76	64.9
11	24.2	44	46.8	77	65.6
12	25.3	45	47.4	78	66.3
13	26.3	46	47.9	79	67.0
14	27.2	47	48.4	80	67.7
15	28.2	48	48.9	81	68.5
16	29.1	49	49.5	82	69.3
17	29.9	50	50.0	83	70.1
18	30.7	51	50.5	84	70.9
19	31.5	52	51.1	85	71.8
20	32.3	53	51.6	86	72.8
21	33.0	54	52.1	87	73.7
22	33.7	55	52.6	88	74.7
23	34.4	56	53.2	89	75.8
24	35.1	57	53.7	90	77.0
25	35.8	58	54.2	91	78.2
26	36.5	59	54.8	92	79.6
27	37.1	60	55.3	93	81.1
28	37.7	61	55.9	94	82.7
29	38.3	62	56.4	95	84.6
30	39.0	63	57.0	96	86.9
31	39.6	64	57.5	97	89.6
32	40.1	65	58.1	98	93.3
33	40.7	66	58.7	99	99.0

* Reprinted from *Resource Handbook on Performance Assessment and Measurement: A Tool for Students, Practitioners, and Policymakers* by Patricia Wheeler and Geneva D. Haertel, 1993, p. 104.

Test Data for Grade 5, 1994-95

Student	Level	Form	Test Date	Raw Score	Scaled Score	Percentile Rank	NCE	Change
Lynn	I	A	10/17/94	29	_____	_____	_____	_____
	II	B	4/18/95	26	_____	_____	_____	_____
Alex	I	A	10/17/94	18	_____	_____	_____	_____
	I	B	4/18/95	23	_____	_____	_____	_____
Kim	II	A	10/17/94	21	_____	_____	_____	_____
	II	B	4/18/95	24	_____	_____	_____	_____
Chris	I	B	10/19/94	31	_____	_____	_____	_____
	II	B	4/18/95	26	_____	_____	_____	_____
Shawn	II	A	10/17/94	14	_____	_____	_____	_____
	I	B	4/18/95	27	_____	_____	_____	_____
Pat	I	A	10/17/94	17	_____	_____	_____	_____
	I	B	4/18/95	18	_____	_____	_____	_____
Tracy	I	B	10/19/94	30	_____	_____	_____	_____
	II	A	4/20/95	24	_____	_____	_____	_____
Lee	II	B	10/19/94	20	_____	_____	_____	_____
	II	A	4/20/95	absent	_____	_____	_____	_____
Ty	I	B	10/19/94	21	_____	_____	_____	_____
	I	A	4/20/95	28	_____	_____	_____	_____
Val	II	A	10/17/94	13	_____	_____	_____	_____
	I	A	4/20/95	21	_____	_____	_____	_____
Kelly	II	A	10/17/94	19	_____	_____	_____	_____
	II	B	4/18/95	21	_____	_____	_____	_____
Summary								
Pre test				***	_____	***	_____	_____
Post test				***	_____	***	_____	_____

*** Do NOT average raw scores or percentile ranks.

Test Data for Grade 5, 1994-95

Student	Level	Form	Test Date	Raw Score	Scaled Score	Percentile Rank	NCE	Change
Lynn	I	A	10/17/94	29	793	35	41.9	
	II	B	4/18/95	26	798	33	40.7	-1.2
Alex	I	A	10/17/94	18	775	11	24.2	
	I	B	4/18/95	23	782	16	29.1	+4.9
Kim	II	A	10/17/94	21	788	27	37.1	
	II	B	4/18/95	24	794	30	39.0	+1.9
Chris	I	B	10/19/94	31	792	33	40.7	
	II	B	4/18/95	26	798	33	40.7	0.0
Shawn	II	A	10/17/94	14	776	13	26.3	
	I	B	4/18/95	27	787	22	33.7	+7.4
Pat	I	A	10/17/94	17	773	10	23.0	
	I	B	4/18/95	18	774	08	20.4	-2.6
Tracy	I	B	10/19/94	30	791	31	39.6	
	II	A	4/20/95	24	793	29	38.3	-1.3
Lee	II	B	10/19/94	20*	788	27	37.1	
	II	A	4/20/95	absent				
Ty	I	B	10/19/94	21	779	16	29.1	
	I	A	4/20/95	28	791	28	37.7	+8.6
Val	II	A	10/17/94	13	774	11	24.2	
	I	A	4/20/95	21	780	13	26.3	+2.1
Kelly	II	A	10/17/94	19	784	21	33.0	
	II	B	4/18/95	21	790	27	37.1	+4.1
Summary								
Pre test				***	782.5	***	31.9	
Post test				***	788.7	***	34.3	+2.4

* Exclude from analysis
since no post-test score.

*** Do NOT average raw
scores or percentile ranks.

References

Gabriel, Roy; & Evans, Warren. (undated). *Tests that fit: An overview of functional-level testing*. Portland, OR: Northwest Regional Educational Laboratory, Title I Technical Assistance Center.

Ingebo, George. (1994). *A study of the accuracy of test measures in functionally administered vs. grade-level administered tests*. Paper presented at the 1994 annual meeting of the California Educational Research Association, San Diego, California.

Wheeler, Patricia; & Haertel, Geneva D. (1993). *Resource handbook on performance assessment and measurement: A tool for students, practitioners, and policymakers*. Chico, CA: The Owl Press. (ERIC Document Reproduction Service No. ED 367 686)